OPEN ACCESS

University of BRISTOL

Ding, Y., & Redmill, DW. (2005). *Direct virtual viewpoint synthesis from multiple viewpoints*. 1045 - 1048. https://doi.org/10.1109/ICIP.2005.1529933

Peer reviewed version

Link to published version (if available):
10.1109/ICIP.2005.1529933

Link to publication record in Explore Bristol Research
PDF-document

University of Bristol - Explore Bristol Research
General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/

Direct Virtual Viewpoint Synthesis from Multiple Viewpoints

Yi Ding
Machine Vision Laboratory, CIMMS,
University of the West of England
Bristol, BS16 1QY, UK
E-mail: yi3.ding@uwe.ac.uk

David Redmill
Center for Communications Research
University of Bristol
Bristol, BS8 1UB, UK
E-mail: david.redmill@bristol.ac.uk

Abstract— **This paper presents a novel approach for synthesizing intermediate or Virtual Viewpoints (VVs) of a 3D scene based on information from a number of known Reference Viewpoints (RVs). The proposed approach directly estimates the pixel value (and corresponding depth) for each pixel in the VV. This is contrast to the more traditional 2 stage approach of firstly building a full 3D or 2.5D model for the scene and then synthesising the desired VV. The potential advantage of this approach is that it works directly with the target virtual view and is hopefully less susceptible to the propagation of errors from the depth estimation stage to the interpolation stage.**

Keywords-direct viewpoint synthesis; depth estimation

I. INTRODUCTION

For many modern applications including special effects for film, TV as well as surveillance and other applications, it is often desirable to generate a high quality **V**irtual **V**iew (**VV**) of a 3D scene from a viewpoint for which no direct information is available. The VV is typically synthesised from information from a number of known **R**eference **V**iews (**RV**s). We shall assume that both intrinsic and extrinsic camera parameters are known for both the VV and RVs. Although these parameters may not be accurately known, they can be derived using various camera calibration algorithms such as those in Hartley and Zisserman [3]. Given these camera parameters, and the position (u,v) and depth z of any pixel in the ith view, it is possible to both project this pixel to a corresponding point in the jth view.

$$\begin{pmatrix} u_j.z_j \\ v_j.z_j \\ z_j \\ 1 \end{pmatrix} = \mathbf{A}_{ij} \begin{pmatrix} u_i.z_i \\ v_i.z_i \\ z_i \\ 1 \end{pmatrix}$$

Providing we can assume that the 3D scene comprises opaque objects, and that any visible point will have similar colour in all viewpoints, it is in principal possible to estimate the 3D scene geometry by matching points in different RVs. While complete 3D scene geometry is useful for many computer vision purposes, it is common for view synthesis purposes to estimate a 2.5D description comprising a depth or disparity estimate for all pixels in each of the RVs. Once the geometry is known, it is then possible to synthesise a virtual view [5]. The first disadvantage of this 2-stage approach is that any errors in the depth estimation stage can propagate via the synthesis stage to cause more significant errors in the desired interpolated view. The second disadvantage is the disparity of every pixel in the RVs estimated irrespective of whether it is visible in the VV. There are two main challenges involved in depth estimation. The first is that of occlusions where portions of the scene may not be visible in some of the RVs because they are occluded by (behind) other objects. For these regions it is unreasonable to expect pixel values to match. The 2nd main problem is that of un-textured regions, which can result in multiple matches at incorrect depth values. In recent years, some authors including Grammalidis et al [1] and Kang et al [3] have observed that using more than two images can dramatically improve the quality of the construction at the expense of increased semi-occluded regions (pixels visible in some but not all images) as the basic pair of stereo views has weakness of dealing with occlusions. In general, this is true for any viewpoint from which additional cameras could be used to disambiguate the possible depth interpretation except some special regions e.g. un-textured regions.

II. DIRECT VIEWPOINT SYNTHESIS

In this paper, we present a novel approach based on directly estimating the depth and corresponding pixel values for the desired VV. Our novel direct viewpoint synthesis bears resemblance with Ng, et al [5] which consists of "Search", "Match" and "Render". Rather than matching in range space, we propose an approach to search in space but match in image space (epipolar scanlines).Tackling the problem in a single stage, has the potential to avoid the problems of inaccurate depth estimates propagating during synthesis. The proposed algorithm involves taking each pixel in the VV and searching along the depth *z* direction. For each candidate depth, we find the corresponding projected pixel values in each of the RVs. The estimated value of depth is then determined as the one which gives the 'best' match between projected pixel values. Having determined the depth, the pixel value can be estimated from the projected pixel values in the RVs. Another way to view this is to consider the pixel in the VV as corresponding to an epipolar line in each of the RVs. We then search along these epipolar lines for a 'best match'.

0-7803-9134-9/05/$20.00 ©2005 IEEE

Authorized licensed use limited to: UNIVERSITY OF BRISTOL. Downloaded on October 1, 2009 at 04:29 from IEEE Xplore. Restrictions apply.

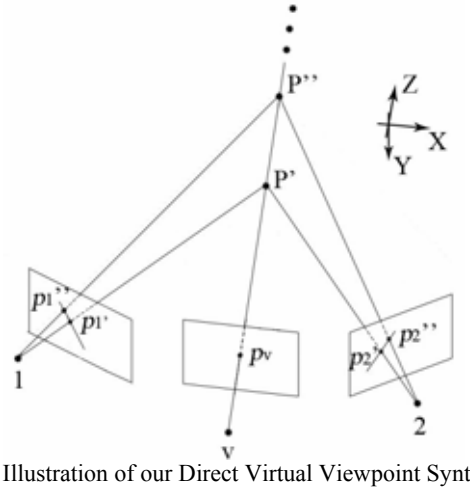

Figure 1: Illustration of our Direct Virtual Viewpoint Synthesis.

Figure 1 shows an illustration of the algorithm for pixel p_v in Virtual Viewpoint v. Searching along the depth direction $P' \rightarrow P''$ maps to the epipolar lines $p_1' \rightarrow p_1''$ in RV_1 and $p_2' \rightarrow p_2''$ in RV_2.

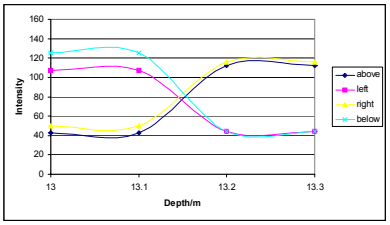

Figure 2: Search process for an example pixel.

Figure 2 illustrates the process for an example pixel in the VV with 4 RVs. The 'best' match is at approximately 13.15m with approximate synthesised pixel intensity of 80. This figure also illustrates that in order to get a good match, we will typically have to search at sub-pixel accuracy [1] and use bi-linear interpolation [6].

In order to utilise this algorithm, we need to formalise both what we mean by 'best' match, and the predicted pixel value. The simplest measure of we can use is to minimise the standard deviation or variance of the projected pixel values, and set the predicted pixel value to the mean of these projected pixel values. We shall refer to this combination of dissimilarity measure and prediction as **M**ean and **S**tandard **D**eviation (**MSD**).

III. Experimental Results

The proposed algorithm has been tested using a set of 4 parallel RVs which are taken from 0.1m to left, right, above and below, the target VV. For comparison purposes, we will compare the synthesised VV to a 'ground truth' value taken from a 5th camera. The cameras all have a focal length of 593.5 pixels, and the minimum and maximum depths for the search process are set to 4.5m and 14m respectively. These depths are chosen from prior knowledge of the scene content to limit the search process. Figures 4a and 5a show the synthesised view and depth map achieved using the proposed MSD algorithm. Figure 6a shows the error compared to the ground truth image. The total **M**ean **S**quared **E**rror (**MSE**) is found to be 44.72. Figure 7a shows the minimum standard deviation (or cost) map.

Looking at the error image (Figure 6a), we can see that the majority of the error occurs around the edges of objects (e.g. lamp and sculpture). Comparing with the depth map (Figure 5a) we see that the majority of the error is associated with large changes in depth. These correspond to regions where the desired data for the VV is occluded in some of the RVs. Comparing with the cost map (Figure 7a) we can also see that the majority of error occurs in regions where the minimum cost was relatively high, i.e. places where we know there is no good match.

Looking at the depth map (Figure 5a), we see three main classes of error. Firstly we have errors at object edges resulting from occlusions (discussed above). Secondly there is significant impulsive noise spread across the whole image. Thirdly there are significant errors in un-textured regions e.g. the shadow under the table and the wall above the whiteboard.

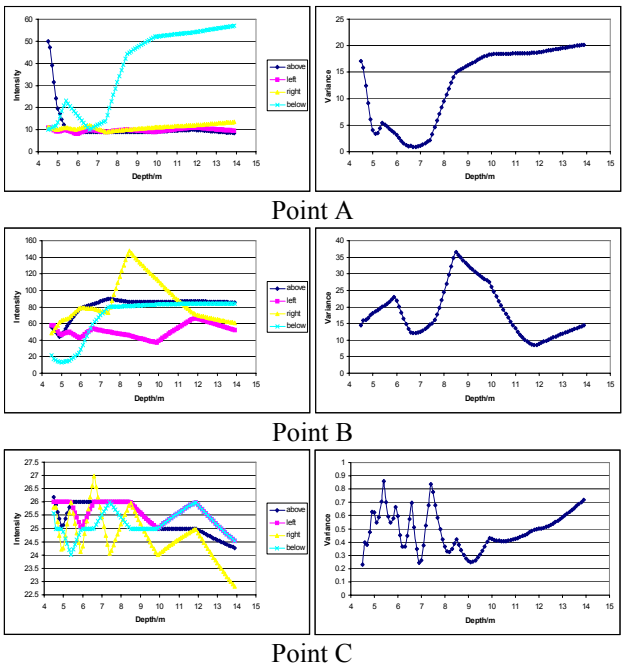

Figure 3: Snapshots of 3 example points showing the projected pixel intensities (left) and standard deviation (right).

To illustrate the algorithms performance more clearly, Figure 3 shows the projected pixel values and standard deviation for 3 different points in the image. At point A we have a single clear global minimum standard deviation at about 6.75m. The clear global minimum implies a high confidence and likelihood of a good synthesis result. At point B, the minimum standard deviation is at about 11.8m. However, the value of this minimum is relatively large, and there are other comparable local minima at 6.8m and 4.5m. The high standard deviation and presence of other comparable local minima at different depths implies that there is no good match and thus we can have low confidence in the synthesis result. This is probably a result of occlusions. At first sight it looks like point C suffers from a similar problem of local minima. However, looking at the values of the standard deviation, we see that they are low for the whole depth range. This is because point C is in

Authorized licensed use limited to: UNIVERSITY OF BRISTOL. Downloaded on October 1, 2009 at 04:29 from IEEE Xplore. Restrictions apply.

an un-textured region. As a result we have low confidence in the depth estimate, although the synthesis result is reasonably good since the synthesised intensity is largely independent of estimated depth.

A. Depth refinement

One reason for the noisy depth map is that the matching is based on individual pixels rather than local regions. While this has advantages in regions with non-uniform depth, it can lead to localised false matches. A simple way to improve the depth map, is to post-filter it with a noise removal filter. In order to preserve sharp discontinuities at object boundaries, it is preferable to use a non-linear filter. A variety of morphological and median filters [7] were tried, with a 3x3 median filter giving the best results. Figures 4b, 5b, and 6b show the synthesised result, depth map and error image after applying a 3x3 median filter to the depth map. The MSE has been reduced from 44.72 to 32.63. The reduction in noise is particularly apparent in the depth map which is visibly better. Note that since this is post-filtering operation, the cost map is no-longer relevant and is thus not shown.

B. New cost function

An alternative to post-filtering the depth map is to use a more sophisticated cost function encompassing both matching local regions and incorporating a smoothness factor. Similar to [4], our cost function for pixel at u,v with depth z is defined as:

$$E(u,v,z) = \sum_{\delta_u,\delta_v} E_{SD}(u+\delta_u, v+\delta_v, z) + E_{smooth}$$

$$E_{smooth}(\eta) = \begin{cases} \gamma(1-\cos(\frac{\pi|\eta|}{2\delta})) & |\eta| \le \delta \\ \gamma & |\eta| > \delta \end{cases}$$

where the constant γ controls the strength of smoothness and η is the depth gradient between the current processing pixel and neighbouring processed pixels in both horizontal and vertical directions. The smoothness term is chosen to encourage smoothness where discontinuities are small $<\delta$ (within an object) and avoid over-smoothing where discontinuities are large $>\delta$ (across object boundaries).

Figures 4c, 5c, 6c and 7c show the synthesised result, depth map, error image and cost map for the proposed direct synthesis algorithm using the proposed cost function with a 3x3 window, $\gamma = 10$ and $\delta = 2$. The MSE is 34.45. While this is slightly more than the result achieved by using MSD with a median post-filter of the depth map, the synthesis does exhibit some advantages, particularly in regions around narrow foreground objects such as the lamp support. The execution times are 33s, 36s and 82s for a) MSD, b) MSD with median filter and c) improved cost function respectively on a PIII 933MHz computer.

IV. Conclusions

This paper has presented a novel direct approach for generating arbitrary Virtual Views (VVs) from a set of known Reference Views (RVs). Although our algorithms are tested using parallel camera setup and just for one intermediate VV, they can be used with an arbitrary camera setup and VV position(s) since no specific geometrical assumptions have been used.

Results are presented which demonstrate the effectiveness of the proposed techniques. It should be noted that for the problem tackled using just 4 reference views, traditional depth estimation techniques would have significant difficulties, since all of the RVs contain some data which is occluded in all the other RVs. This would lead to significant problems in depth estimation which would manifest themselves as significant problems in the synthesis.

The proposed method is relatively efficient for synthesising a single VV since depth estimation is only performed once for the VV rather than many times for each of the RVs. However, if we wish to synthesise many difference VVs, then complexity may be a serious issue.

The proposed approach is a novel approach which could benefit from further research to improve the cost function. It could also be easily extended to problems with more RVs. A further possible refinement would be to utilise colour information rather than just luminance. Further work is also needed to compare the proposed approach with a traditional 2-stage approach.

References

[1] S. Birchfield, et al, "Depth Discontinuities by Pixel-to-Pixel Stereo", *Sixth International Conference on Computer Vision*, Bombay, India, pp. 1073-1080, January 1998.

[2] N. Grammalidis, et al, "Disparity and Occlusion Estimation in Multiocular Systems and Their Coding for the Communication of Multiview Image Sequences", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 8, No. 3, pp. 328-344, June 1998.

[3] R. Hartely, and A. Zisserman, Multiple View Geometry in Computer vision, Cambridge University Press, 2000.

[4] S. –B. Kang, et al, "Handling Occlusions in Dense Multi-view Stereo", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, I-103-I-110, December 2001.

[5] K. C. Ng, et al, "Generalized Multiple Baseline Stereo and Direct Virtual View Synthesis Using Range-Space Search, Match, and Render", *International Journal of Computer Vision*, *Special Issue on Multicamera Stereo*, Vol. 47, Numbers 1/2/3, pp. 131-147, April-June 2002.

[6] D. Scharstein, "View synthesis using stereo vision", *Lecture notes in computer science*, Springer, October 1999.

[7] M. Sonka, and V. Hlavac and R. Boyle, *Image Processing, Analysis, and Machine Vision*, Brooks/Cole Publishing Company, 1999.

Authorized licensed use limited to: UNIVERSITY OF BRISTOL. Downloaded on October 1, 2009 at 04:29 from IEEE Xplore. Restrictions apply.

Figure 4: Synthesised view using: a) MSD, b) MSD + median filter of depth map, c) Improved cost function

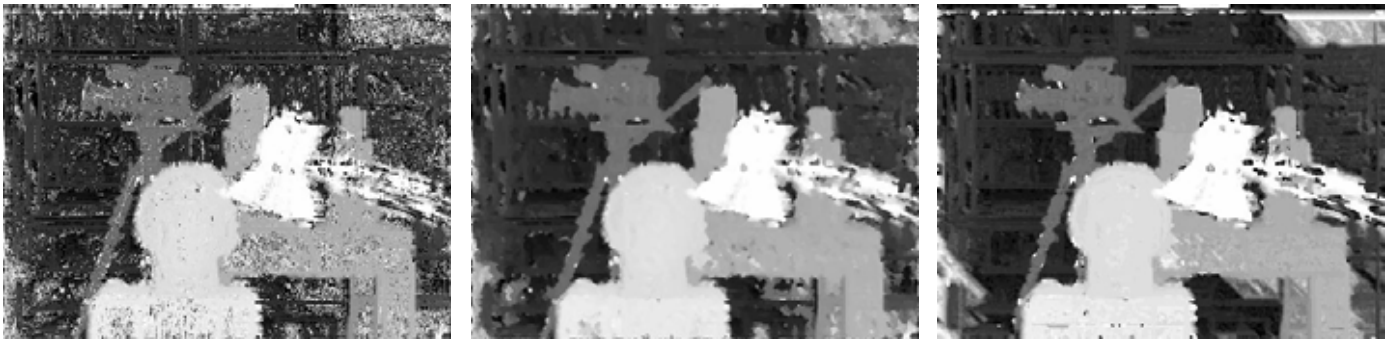

Figure 5: Synthesised depth map using: a) MSD, b) MSD + median filter of depth map, c) Improved cost function

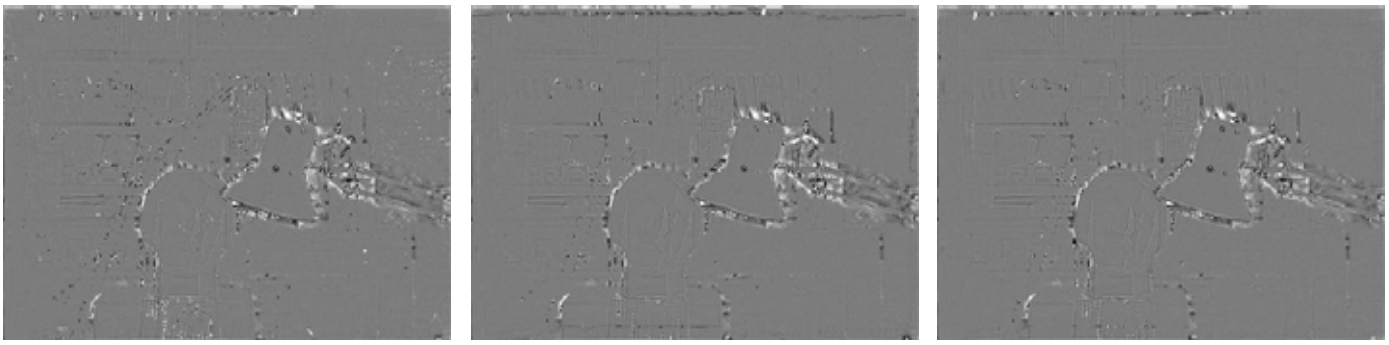

Figure 6: Error image using: a) MSD, b) MSD + median filter of depth map, c) Improved cost function. MSE = 44.72, 32.63, and 34.45 respectively. Note that in order to exclude boundary effects, these MSE figures exclude pixels within 5 pixels of the image edges.

Figure 7: Cost map using: a) MSD, c) Improved cost function. Note that the median filter of the depth map doesn't alter the cost map which is the same as MSD.

Authorized licensed use limited to: UNIVERSITY OF BRISTOL. Downloaded on October 1, 2009 at 04:29 from IEEE Xplore. Restrictions apply.